Recording and timing vocal responses in online experimentation

Katrina Kechun Li†, Julia Schwarz†, Jasper Hong Sim, Yixin Zhang, Elizabeth Buchanan-Worster, Brechtje Post, Kirsty McDougall

University of Cambridge

†These authors have contributed equally to this work and share first authorship

kl502@cam.ac.uk, js2275@cam.ac.uk

Abstract

Cued shadowing is a psycholinguistic task that captures the response speed and accuracy of participants' vocal repetition of target words. Due to its simplicity, the paradigm is widely used as a naturalistic measure of speech processing. While the COVID-19 pandemic has driven the adaptation of many lab-based experiments to internet-based data collection, cued shadowing is not straightforward to adapt due to various challenges, including the precision of timing, efficient extraction of response latencies, and control over data quality. The current paper presents solutions to these challenges and describes the methodology for conducting cued shadowing of audio-video stimuli online with children and adults. The performance of two (semi-)automatic speech onset detection tools and two experimental designs are evaluated. The technique developed enables millisecond precision in response time measurement and has great potential for the inclusion of minority and hard-to-reach communities in future speech perception and production research.

Index Terms: online experimentation, cued shadowing, remote data collection, reaction time

1. Introduction

A growing number of lab-based psycholinguistic experiments have been adapted to an internet-based format, including designs that involve the recording of spoken responses. Recent work has demonstrated that voice recordings can be collected through various modes in non-laboratory settings [1, 2], and can yield accurate phonetic measurements of duration and f0 [3, 4], as well as reaction time measurements (i.e., response latencies) of comparable precision to lab-based experiments when collected with a picture-naming task [5, 6]. A related popular technique, cued shadowing, poses several technical challenges for online data collection and as such has not yet been adapted to internet-based experimentation. The current paper addresses these challenges by presenting a novel design for internet-based audio-visual cued shadowing with children and adults, and evaluates different methods for collecting and measuring response latencies.

1.1. Challenges of internet-based cued shadowing

In the cued shadowing paradigm (also known as 'auditory word repetition'), participants orally repeat auditorily presented target words. Cued shadowing does not require literacy or metalinguistic knowledge and as such can be used as a naturalistic measure of listening effort, i.e., of the resources required by a listener to meet the cognitive demands of processing speech accurately and efficiently [7, 8]. Moreover, internet-based cued shadowing is time- and cost-effective and enhances ecological validity by allowing participants to complete the research in familiar surroundings. It is therefore well-suited to children and adults as well as minority and hard-to-reach communities.

A major challenge for internet-based cued shadowing is the precise time-locking of voice recordings to the presentation of audio-visual stimuli. The continued development of online experiment tools in recent years has increased the reliability of reaction time measurements [9, 10]. Several well-established effects found in lab-based psycholinguistic experiments have been replicated using a variety of internet-based experiment platforms [11, 12, 13]. Many of the paradigms involve the presentation of static stimuli and reaction time measurements of keyboard responses. However, the complications involved in the presentation and timing of audio-visual stimuli in relation to the use of voice recording functions have yet to be addressed.

Another challenge is that extracting reaction times from online collected vocal responses is less straightforward than collecting reaction times from keyboard responses. Although a number of experiment platforms include a voice recording function [5, 6], no internet-based experiment tools feature a voice key, i.e., specialist software for automatically detecting the onset of vocal responses. Thus, the development of robust methods for extracting response latencies efficiently from online recordings is required.

Furthermore, concerns have been raised about the quality of data collected online. Since participants cannot be monitored closely by the researcher, distraction, background noise, and a lack of commitment to the task can negatively impact the data quality. With respect to cued shadowing, various technical issues can occur, such as interruptions in playing media files and suboptimal recording quality, thereby influencing measurement accuracy.

1.2. The current paper

The current paper presents and evaluates the methodology for collecting response latencies with an internet-based cued shadowing paradigm in the context of a research study on the processing of face mask speech (for the full findings, see [14]). The aim of the study was to measure the extent to which children (aged 8-12) and adults (aged 20-60) experience processing difficulties with face mask speech. The experiment was implemented with Gorilla experiment builder [15]. Participants were presented with audio-video sentence stimuli and asked to repeat the last word of each sentence as quickly as possible. Stimuli were manipulated (1) acoustically by presenting the audio signal produced with/without a mask, (2) visually by displaying the speaker with/without mask, and (3) semantically by varying the predictability of the sentence-final target words (cloze probability [16, 17]). For example, the

target word ‘cake’ was embedded in the high Cloze Probability context (“For your birthday I baked this cake”) and in the Low Probability context (“Tom wants to know about this cake”).

2. Methodology

2.1. Experimental design

2.1.1. Audio-video stimuli

The stimuli consisted of 120 target words embedded in 240 English sentences with high and low predictability, which were adapted from [17]. Carrier sentences contained five to eight words, and all target words were monosyllabic nouns starting with a consonant. Stimuli were spoken by a female native English speaker with Standard Southern British English accent, with and without a cloth face mask. Audio recordings (sampling rate of 44.1 kHz at 16 bits) and video recordings (1920-by-1080 resolution at 50 fps) were made simultaneously in a sound-attenuated recording booth, and subsequently synchronised in Final Cut Pro (version 10.5.2). Each video recording (with/without face mask) was paired with the audio recordings (with/without face mask). The final audio-video stimuli were saved in mp4 format (H.264 video codec, AAC audio codec).

2.1.2. Trial design

Each trial began with a 250 ms fixation cross, followed by the audio-video presentation of the stimulus. Participants watched the audio-video stimulus and then orally repeated the sentence-final target word before moving on to the next trial. A progress bar and the correct answer were displayed on the screen for 1500 ms before automatically continuing to the next trial.

In order to capture response times accurately, three essential features that deviate from typical trial designs for picture naming were implemented [5, 6]. First, instead of starting the voice recording function after presenting each stimulus, the recording automatically started at the beginning of each trial in order to record the stimulus together with the participant’s response. This feature was implemented because participants sometimes gave a response before the end of the trial display, and because there can be a delay between the request for the microphone to begin recording and when the microphone actually starts recording. Secondly, a single-peak acoustic signal of 500 Hz (a ‘beep’) was inserted in the audio-track of the stimuli before the start of the presented sentence. The beep was created in *Praat* using the formula in (1).

$$0.7e^{\left(-0.5\left(\frac{x}{0.01}\right)^2\right)}sin(2\pi 500x) \quad (1)$$

The beep not only indicated to the participants that the stimulus was going to be played, but more importantly also served as an anchor for calculating response latencies. The beep was captured more reliably than the spoken part of the stimuli by the participants’ computer microphones and could also be easily identified by the peak of intensity in the waveform during data processing. Thirdly, the display of audio-video stimuli was started manually by participants since automatically playing videos (i.e., the auto-play feature) is not enabled in all browsers and might introduce timing inaccuracies.

Including the stimuli within the recording file allowed precise calculations of response latencies. To calculate response latencies, the total duration from the beep to speech onset was extracted (X1, Figure 1). The duration of each sentence stimulus was measured separately (X2) and then subtracted from the total trial duration to derive response latencies from stimulus offset to speech onset. In addition, recording the complete trial sequence enabled more effective monitoring of data quality, as the recordings gave an indication as to whether there were any technical issues or distractions during the experiment.

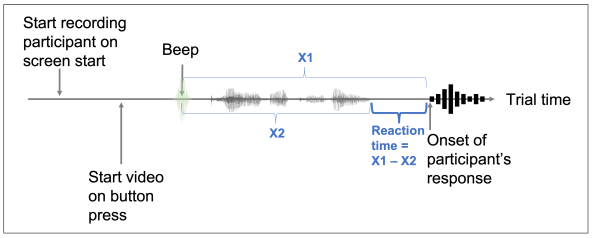

Figure 1: *Trial design for capturing response latencies. X1: Duration of trial recording from the beep to the response onset. X2: Duration of stimulus from the beep to the end of the presented sentence.*

A possible alternative to this design would have been to rely on the timing metrics of video events for the stimulus onset provided by the experiment platform; however, the robustness of these metrics has not yet been tested. Therefore, the current design of measuring response latencies from an embedded sound anchor will be compared with response latencies derived from the video start metric provided by Gorilla in Section 3.3.2.

2.1.3. Procedure

Participants performed a speaker and microphone test, and then were presented with written instructions and an animated demonstration of the task. Participants were instructed to use loudspeakers so that the output of the audio stimuli, especially the beep, would be captured in the recording. They were instructed to say ‘I don’t know’ during the task if they did not know the correct answer, as empty recordings are hard to interpret. The children’s version of the experiment introduced a parrot named Polly and asked the children to ‘teach’ Polly new words. The experiment started with 12 practice items and a reminder of the instructions. 120 trials were equally distributed across four blocks. Finally, a post-experiment questionnaire asked participants to report any technical difficulties that occurred during the experiment.

2.2. Data Processing

2.2.1. Speech onset detection

One of the tools developed for automatic speech onset detection, Chronset [18], has recently been applied for marking responses in both web-based and lab-based picture naming tasks [5, 19, 20]. Chronset is an easy-to-use online tool that detects speech onsets by relying on multiple acoustic features. However, Chronset is unable to identify the onset of a participant’s response if the recording file to be analysed includes additional speech sounds as is the case in the current design. Although most of Chronset’s measurement errors lie within 50 ms [18], this variation could obscure small psycho-linguistic effects. Therefore, a data processing pipeline was developed using *Praat* [21] (scripts available on OSF [22]) to derive the duration from the beep to the speech onset in each recording file (X1, Figure 1). The pipeline made use of Chronset and an additional, intensity-based tool to facilitate accurate speech onset detection:

1. For each trial, an interval containing the beep and an interval with the participant’s response were marked in a *Praat* textgrid.

2. In the intervals that contained the beep, the time point of the maximum intensity was extracted with a *Praat* script.

3. The intervals containing participants' responses were separated from the full recording file, and 50 ms of silence was added at the beginning and end of each file to allow correct functioning of the speech onset detection tools. Speech onsets were detected with Chronset and the custom *Praat* script. The latter marked the response onset based on an intensity threshold that could be manually adjusted to each participant.

4. The automatically identified speech onsets were manually corrected by one of three trained phoneticians (henceforth 'raters') with partial cross-checking. To facilitate this process, a *Praat* script was developed to zoom the view window automatically to +/- 50 ms of the estimated speech onset and allow raters to accept or correct the speech onset marking. The corrected speech onset was then automatically aligned to the nearest zero crossing.

2.2.2. Criteria for manual speech onset correction

The manual corrections of response onsets followed commonly used phonetic criteria that all raters agreed upon before taking measurements, thereby minimising any differences in phonetic judgement. For sonorants (e.g., [r], [m]), the onset was defined as the first upward-going zero crossing of the regular sinusoidal curve on the waveform. For plosives (e.g., [p], [p^h]), the onset was measured from the start of the spike that indicated a burst, in accordance with voice onset time measurements [23]. Fricative onset (e.g., [f], [ʃ]) was defined as the point at which high frequency energy first appeared on the spectrogram and/or the point at which the number of zero crossings rapidly increased [24]. When preceding voicing was present in voiced fricatives or voiced plosives, e.g., pre-nasalisation or pre-voicing, the start of voicing was considered the speech onset.

2.2.3. Participant Screening

Participants were screened through several means, which addressed most concerns over the quality of online collected data: (1) the post-experiment questionnaire in which participants self-reported whether they had encountered any technical problems with the video, audio, or voice recordings (Section 2.1.3); (2) Gorilla's experiment platform metrics, containing information about video loading delays and recording errors; (3) screening of the voice recording files by the researchers for unwanted noise, distractions, or suboptimal quality.

3. Results and Evaluations of Methods

3.1. Data quality

Out of 78 adults and 67 children who completed the experiment between July and September 2021, 21% of the participants were excluded for not fulfilling the recruitment criteria or not following the instructions. 30% of the participants were excluded due to the occurrence of one or more technical problems with the audio-video display or the voice recordings. This left a participant group of 40 adults and 26 children. The relatively large video size (10 MB) was likely to be the main cause of video playback problems, suggesting that stimuli files should be compressed to a smaller size. Furthermore, word-initial fricatives were less reliably recorded and thus difficult to measure, so researchers may want to avoid target words with a fricative onset in future online experiments. Individual trials were also discarded if the responses were of poor recording quality or if the speech onset could not be identified reliably (0.50% of 5653 trials, from 26 children and 26 adults).

3.2. Results of response latencies

A full report of the results can be found in the associated paper on face mask speech processing [14]. Response times of all accurate responses were analysed with linear mixed-effects models. The results revealed both small-sized effects (under 50 ms) and medium-sized effects (under 100 ms) comparable to lab-based experiments. The model showed significant main effects of acoustic mask, visual mask, and cloze probability on response latencies in the expected direction: On average, responses to acoustically masked speech were 30 ms slower than unmasked speech (b = -12.64, 95% CI [-18.18, -7.09], $p <$ 0.001), responses to a video of the speaker with a visual mask were 30 ms slower than the same speaker without a mask (b = -13.11, 95% CI [-18.73, -7.48], $p < 0.001$), and responses to low cloze probability targets were 79 ms slower than high probability targets (b = -40.89, 95% CI [-53.88, -27.90], $p <$ 0.001).

3.3. Evaluation

The manually corrected response latencies of the present design were compared to the response times derived from two automatic speech detection methods without corrections: Chronset and an intensity-based *Praat* script (Section 3.3.1.).

The design of the present study was also compared to an alternative design that relies on the video's 'timeupdate event' instead of an inserted sound anchor in the audio track of the stimuli (Section 3.3.2.). The 'timeupdate event' is provided by Gorilla and is supposed to capture the precise start of the presented video stimuli. Provided that this metric is accurate, the calculation of response latencies would no longer require a sound anchor, thereby significantly simplifying the design. Instead, response latencies could be derived by subtracting the video start time and stimulus duration (cf. Figure 2).

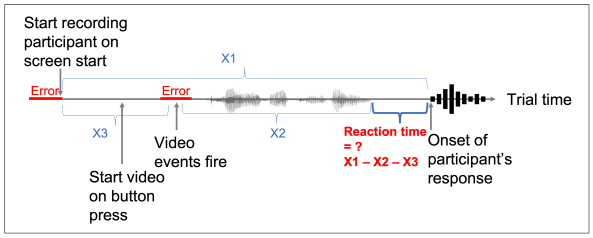

Figure 2: *Alternative design for capturing response times based on video metric. X1: Duration from the start of trial recording to speech onset. X2: Duration from file start to sentence end. X3: Video event start minus recording start as captured by Gorilla.*

Two measures were used to evaluate performance [18]: (1) absolute-difference scores, reflecting the absolute difference between response times derived from manual vs. automatic speech onset detection, and (2) regression fit, with manual measurements as the dependent variable. The regression fit quantifies the relationship between automatic estimates and manual ratings based on fit (R^2) and regression residuals (the unobserved error that can be attributed to either the manual or the automatic scores). All comparisons are visualised in Figures 3 and 4.

3.3.1. *Comparison of automated and manual measurements of speech onset*

With respect to absolute-difference scores, 75% of the measurements from Chronset differed by 50 ms or less from the manual measurements, while 28% differed by 10 ms or less. In comparison, over 83% of the intensity-based measurements deviated by 50 ms or less from the original measurements, and 46% by 10 ms or less. While the performance of the two tools was comparable for affricates, plosives, and sonorants, there was a large discrepancy in performance for fricatives. While 77% of intensity-based fricative measurements deviated from manual measurements by 50 ms or less, only 63% of Chronset measurements achieved this level of precision.

With respect to regression fits, a strong linear relationship was observed between manual ratings and automatic scores from Chronset ($b = 0.80$, $R^2 = 0.80$, Figure 3 (A)) and between manual ratings and the intensity-based tool ($b = 0.61$, $R^2 = 0.60$, Figure 3 (B)). Figure 4 shows the cumulative density function of the regression error of the intensity-based script (black line) and Chronset (orange line) compared to manual measurements.

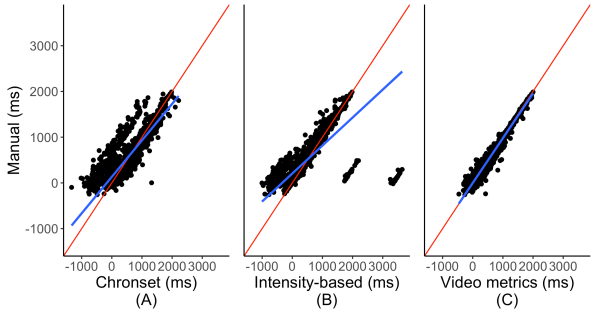

Figure 3: *Comparison of manual and automated measurements for calculating response latencies. The red lines represent the lines of identical correspondence, and the blue lines represent the regression fit.*

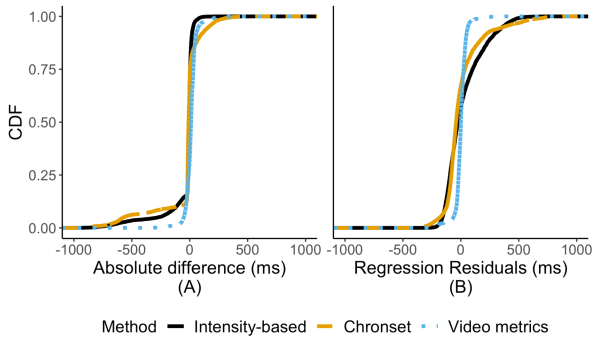

Figure 4: *Empirical cumulative distribution function of the absolute difference of speech onset estimates and regression residuals relative to manual markings.*

Taken together, the intensity-based measurements showed more outliers (i.e., onset estimates strongly deviating from manual measurements) than Chronset measurements, but achieved better precision in terms of absolute-difference after excluding outliers. Because the intensity threshold could be adjusted to each participant, the measurements based on intensity provided a more convenient baseline for manual correction. Chronset measurements yielded a stronger correlation with manual measurements and more concentrated regression errors than the intensity-based estimates. However, the overall performance of Chronset was lower than reported in [18], deviating more strongly from the raters' manual speech onset markings than intensity-based measurements. This could be the case because Chronset relies on different onset marking criteria than those used by the raters or because of the reduced quality of recordings collected online compared to responses collected with professional equipment.

3.3.2. *Comparison of current experiment design to the design using video event metric*

Response latencies calculated with Gorilla's video 'timeupdate event' achieved results close to the original trial design (presented in Section 2.1.2.): 85% of the errors deviated by 50 ms or less from the original measurements, and 23% of the errors deviated by 10 ms or less. The regression model showed few outliers and the coefficient was close to one ($b = 0.98$, $R^2 = 0.98$), indicating an almost perfect correspondence between the two designs (shown in Figures 3 (C) and Figure 4 respectively). This finding provides the first empirical evidence that timing measurements derived from video 'timeupdate event' metrics are highly reliable. Therefore, future cued shadowing designs and online data processing can be simplified. The new design would also allow the use of headphones for better stimulus sound quality, since it does not require the recording of a 'beep' signal through the participants' device. A shortcoming of the method using the 'timeupdate event' metric, however, is that sometimes video files may not be successfully pre-loaded, resulting in inaccurate video playing metrics. In the present evaluation, a further 14% of the data (787 trials) were lost due to inaccurate event metrics. Furthermore, close monitoring of data quality would not have been possible by relying on platform metrics only.

4. Conclusions

The present paper demonstrates that response latencies can be collected accurately with internet-based cued shadowing from children and adults. The trial design presented here, which relied on a sound anchor being recorded together with the participants' vocal responses, provided accurate measurements for response times and enabled effective monitoring of data quality. Effect sizes for small (30 ms) and medium-sized (79 ms) effects were comparable to lab-based studies. The speech onset detection tools (Chronset and an intensity-based *Praat* script) facilitated manual markings of online-collected speech data, but the evaluation revealed that their correct functioning relies on the recording quality of the collected data. Moreover, measuring response latencies from a sound anchor to mark the beginning of a new stimulus aligned almost perfectly with measuring latencies using the video 'timeupdate event' metric provided by the experiment platform. Using 'timeupdate' metric could simplify the online collection of vocal response latencies in the future if trial loss from technical errors can be reduced.

5. Acknowledgements

The authors would like to thank Hugo Caffaratti and the Gorilla Support team for their helpful suggestions on various technical questions. This research was funded by a grant from the Cambridge Language Sciences Incubator Fund and the Isaac Newton Trust.

6. References

[1] N. H. Hilton and A. Leemann, "Editorial: using smartphones to collect linguistic data", *Linguistics Vanguard*, vol. 7, no. s1, Jan. 2021, doi: 10.1515/lingvan-2020-0132.

[2] C. Sanker et al., "(Don't) try this at home! The effects of recording devices and software on phonetic analysis", *Language*, vol. 97, no. 4, pp. e360–e382, 2021, doi: 10.1353/lan.2021.0075.

[3] C. Ge, Y. Xiong, and P. Mok, "How reliable are phonetic data collected remotely? Comparison of recording devices and environments on acoustic measurements", in *Proc. Interspeech*, Aug. 2021, pp. 3984–3988. doi: 10.21437/Interspeech.2021-1122.

[4] C. Zhang, K. Jepson, G. Lohfink, and A. Arvaniti, "Comparing acoustic analyses of speech data collected remotely", *J. Acoust. Soc. Am.*, vol. 149, no. 6, pp. 3910–3916, Jun. 2021, doi: 10.1121/10.0005132.

[5] A. Vogt, R. Hauber, A. K. Kuhlen, and R. A. Rahman, "Internet-based language production research with overt articulation: Proof of concept, challenges, and practical advice", *Behav Res*, Nov. 2021, doi: 10.3758/s13428-021-01686-3.

[6] A. Fairs and K. Strijkers, "Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors", *PLOS ONE*, vol. 16, no. 10, e0258908, Oct. 2021, doi: 10.1371/journal.pone.0258908.

[7] R. Houben, M. van Doorn-Bierman, and W. A. Dreschler, "Using response time to speech as a measure for listening effort", *International Journal of Audiology*, vol. 52, no. 11, pp. 753–761, Nov. 2013, doi: 10.3109/14992027.2013.832415.

[8] J. E. Peelle, "Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior", *Ear & Hearing*, vol. 39, no. 2, pp. 204–214, Mar. 2018, doi: 10.1097/AUD.0000000000000494.

[9] A. Anwyl-Irvine, E. S. Dalmaijer, N. Hodges, and J. K. Evershed, "Realistic precision and accuracy of online experiment platforms, web browsers, and devices", *Behav Res*, vol. 53, no. 4, pp. 1407–1425, Aug. 2021, doi: 10.3758/s13428-020-01501-5.

[10] S. Reimers and N. Stewart, "Auditory presentation and synchronization in Adobe Flash and HTML5/JavaScript Web experiments", *Behav Res*, vol. 48, no. 3, pp. 897–908, Sep. 2016, doi: 10.3758/s13428-016-0758-5.

[11] K. Enochson and J. Culbertson, "Collecting psycholinguistic response time data using Amazon Mechanical Turk", *PLOS ONE*, vol. 10, no. 3, e0116946, Mar. 2015, doi: 10.1371/journal.pone.0116946.

[12] J. Kim, U. Gabriel, and P. Gygax, "Testing the effectiveness of the Internet-based instrument PsyToolkit: A comparison between web-based (PsyToolkit) and lab-based (E-Prime 3.0) measurements of response choice and response time in a complex psycholinguistic task", *PLOS ONE*, vol. 14, no. 9, e0221802, Sep. 2019, doi: 10.1371/journal.pone.0221802.

[13] F. Keller, S. Gunasekharan, N. Mayo, and M. Corley, "Timing accuracy of Web experiments: A case study using the WebExp software package", *Behav Res*, vol. 41, no. 1, pp. 1–12, Feb. 2009, doi: 10.3758/BRM.41.1.12.

[14] J. Schwarz, K. K. Li, J. H. Sim, Y. Zhang, E. Buchanan-Worster, B. Post, J. Gibson, and K. McDougall, "Semantic cues modulate children's and adults' processing of audio-visual face mask speech", *Front. Psychol*., 2022.

[15] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder", *Behav Res*, vol. 52, no. 1, pp. 388–407, Feb. 2020, doi: 10.3758/s13428-019-01237-x.

[16] H. Liu, E. Bates, T. Powell, and B. Wulfeck, "Single-word shadowing and the study of lexical access", *Appl. Psycholinguistics*, vol. 18, no. 2, pp. 157–180, Apr. 1997, doi: 10.1017/S0142716400009954.

[17] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability", *J. Acoust. Soc. Am.*, vol. 61, no. 5, pp. 1337–1351, May 1977, doi: 10.1121/1.381436.

[18] F. Roux, B. C. Armstrong, and M. Carreiras, "Chronset: An automated tool for detecting speech onset", *Behav Res*, vol. 49, no. 5, pp. 1864–1881, Oct. 2017, doi: 10.3758/s13428-016-0830-1.

[19] H. S. Gauvin, M. K. Jonen, J. Choi, K. McMahon, and G. I. de Zubicaray, "No lexical competition without priming: Evidence from the picture–word interference paradigm", *Quart. J. Exp. Psychol.*, vol. 71, no. 12, pp. 2562–2570, Dec. 2018, doi: 10.1177/1747021817747266.

[20] A. de Bruin, A. G. Samuel, and J. A. Duñabeitia, "Voluntary language switching: When and why do bilinguals switch between their languages?", *J. Mem. Lang.*, vol. 103, pp. 28–43, Dec. 2018, doi: 10.1016/j.jml.2018.07.005.

[21] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2019. Accessed: May 26, 2019. [Online]. Available: http://www.praat.org/.

[22] J. Schwarz, K. K. Li, J. H. Sim, Y. Zhang, E. Buchanan-Worster, B. Post, J. Gibson, and K. McDougall, "PerMaSC: Speech perception through masks in school contexts," 13-Apr-2021. [Online]. Available: osf.io/etvdg.

[23] T. Cho and P. Ladefoged, "Variation and universals in VOT: evidence from 18 languages", *J. Phonetics*, vol. 27, no. 2, pp. 207–229, Apr. 1999, doi: 10.1006/jpho.1999.0094.

[24] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of English fricatives", *J. Acoust. Soc. Am*, vol. 108, no. 3, pp. 1252–1263, Sep. 2000, doi: 10.1121/1.1288413.